Optimal Experimental Design for Prediction Based on Push-forward Probability Measures

Butler, T.[b], Jakeman, J.[a,1], Wildey, T.[a,1,*]

[a] *Sandia National Laboratories, Center for Computing Research, Albuquerque, NM 87185, United States*
[b] *University of Colorado Denver, Department of Mathematical and Statistical Sciences*

Abstract

Incorporating experimental data is essential for increasing the credibility of simulation-aided decision making and design. This paper presents a method which uses a computational model to guide the optimal acquisition of experimental data to produce data-informed predictions of quantities of interest (QoI). Many strategies for optimal experimental design (OED) select data that maximize some utility that measures the reduction in uncertainty of uncertain model parameters, for example the expected information gain between prior and posterior distributions of these parameters. In this paper, we seek to maximize the expected information gained from the push-forward of an initial (prior) density to the push-forward of the updated (posterior) density through the parameter-to-prediction map. The formulation presented is based upon the solution of a specific class of stochastic inverse problems which find a probability density that is consistent with the model and the data in the sense that the push-forward of this density through the parameter-to-observable map matches a given density on the observable data. While this stochastic inverse problem forms the mathematical basis for our approach, we develop a one-step algorithm, focused on push-forward probability measures, that leverages inference-for-prediction to by-

*Corresponding author
Email address: tmwilde@sandia.gov (Wildey, T.)

[1] This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

pass constructing the solution to the stochastic inverse problem. A number of numerical results are presented to demonstrate the utility of this *optimal experimental design for prediction* and facilitate comparison of our approach with traditional OED.

Keywords: Uncertainty quantification, optimal experimental design, push-forward measures, stochastic inference

---

1. Introduction

Developing techniques for assimilating experimental data is essential for increasing the credibility of simulation-aided decision making and design since this data can be used to reduce uncertainty and inform model structure. However, when the collection of experimental data is expensive, only a limited amount of experimental data can be obtained. Moreover, different experiments provide different amounts of information about the processes they are helping inform. Consequently, it is crucial to design experiments for which the associated data maximizes the value, or the information content, of each experiment.

The simplest methods for optimal experimental design (OED) employ heuristics, based on concepts such as space-filling and blocking, to select field experiments [13, 15, 27, 29, 31]. These methods can typically be improved upon by utilizing numerical simulation of physical processes as a model for the data likely to be observed and then designing experiments which minimize model uncertainty [6, 7, 8, 18]. When model observables are linear with respect to the model parameters, many popular OED approaches seek designs that reduce uncertainty in the statistical estimates of the model parameters by maximizing some scalar function of the Fisher information matrix, denoted here by M. For example, A-optimal designs maximize the trace of the inverse of M, thereby minimizing the average estimation variance. In contrast, D-optimality maximizes the determinant of M with the goal of minimizing the volume of the confidence ellipsoid around the model parameters. These, and other so-called alphabetic criteria, have been developed in both Bayesian and Frequentist settings [2, 1, 17, 4, 11, 23].

For non-linear models, Bayesian formulations provide a general framework for OED [24, 26, 30, 2, 22, 16, 18]. These approaches assign a prior distribution to the uncertain model parameters and then determine a design by maximizing a utility function over a set of possible designs. Various forms of the utility function are used, but a popular approach is to maximize the expected information gain between the prior and the posterior

distribution. Such approaches are often formulated as an inverse problem wrapped within an optimization procedure. Subsequently, Bayesian formulations tend to be computationally expensive, often requiring thousands or millions of model evaluations. Computationally efficient and scalable approaches for OED have been the focus of many recent efforts [2, 22, 23] and often leverage Laplace approximations and connections with scalable methods for deterministic optimization.

In this paper, we are primarily concerned with designing experiments that maximize the expected information content when making simulation-based predictions. For example, consider a model of the flow of fluid around a set of obstacles used to predict the drop in pressure behind each object. Suppose the locations of the obstacles are unknown so that the pressure drops cannot be measured directly. In this case, we can choose to observe the magnitude of flow velocity at certain predetermined locations. Then, we can use parameter estimation and a simulation model of the observable quantities to infer the location of each object. The resulting parameter uncertainties can then be propagated through a (possibly different) simulation model to obtain the prediction QoI, i.e. the drop in pressure at each obstacle. This process is often referred to as *inference-for-prediction*. In general, the observational model and prediction model need not be the same, however they must both contain the same set of uncertain variables as input parameters. For such situations where prediction is the ultimate objective, experimental designs that target reduction in parameter uncertainty may be inefficient, or worse yet, entirely ineffective. This is due to the fact that data collected may only inform certain directions/regions of the parameter space while the prediction QoI may only exhibit sensitivities to other directions/regions.

The focus of this work is on an *optimal experimental design for prediction* (OED4P) procedure that seeks to improve the predictive capability of a simulation. This goal-oriented approach can significantly reduce the amount of data that needs to be collected to inform a prediction. We can subsequently exploit the low-dimensionality of the data space to reduce the computational complexity of determining the OED4P solution. Several existing OED approaches also focus on uncertainty reductions for prediction. For example, I-optimality minimizes the average prediction variance and G-optimality minimizes the worst-case prediction variance. However typically these criteria focus on reducing uncertainty in predictions of observations that were considered in the design process. In contrast, the approach in this work generates designs that consider uncertainty in QoI possibly unrelated to the observations. The concepts presented here are synergistic with the

ideas developed in [20, 21], which sought to bypass the parameter space in the process of inference–for–prediction, given fixed data. Recently scalable approaches for goal-oriented Bayesian inference have been developed for models parameterized by random variables with Gaussian prior distributions [5] and additive Gaussian noise.

The approach proposed in this paper is fundamentally different from previous OED4P approaches in that it uses a data-consistent framework for formulating the stochastic inversion problem [9] which facilitates the development of a new algorithmic procedure for defining a *data–consistent prediction* (i.e., an updated prediction measure) that entirely bypasses the solution to the stochastic inverse problem. This data-consistent framework is based on a different class of inverse problems than the one typically addressed by traditional Bayesian methods. Specifically, instead of seeking posterior distribution that is conditioned by the data, this framework seeks a probability measure on the model inputs parameters that is consistent with the data in the sense that the corresponding push-forward distribution matches a given target distribution. Such problems arise in many engineering, manufacturing, design and biomedical applications where the variability in observations can be attributed to variability in the model parameters across a population or collection of manufactured components. Despite this significant difference in the problem formulation, there are also some similarities. In particular, both approaches utilize prior (or initial) information about the model input parameters to regularize the respective inverse problems to provide unique and stable solutions. In [9], we discuss these similarities and differences in more detail and provide a simple set of examples to compare the solutions and predictions for each approach. The algorithms described in this work only require a mechanism for generating samples from the initial (prior) distribution and ability to define the push–forward of these samples onto the joint observable/prediction space. An additional feature of the algorithms is that all numerical integrals and approximations occur exclusively in either the observation space or the prediction space; never in the full joint space or in the parameter space.

The main contributions of this paper are:

- a definition of the inference-for-prediction problem based on push-forward measures;
- the formulation of OED for prediction based entirely on these push-forward measures;
- an intuition-building discussion of the special case of linear maps and

Gaussian distributions;

- a discussion of the assumptions made in this paper to achieve a scalable formulation for computing and optimizing the expected information gained for prediction;
- several numerical examples highlighting the utility of OED for prediction and the differences between the designs it selects versus those selected using traditional OED.

For the sake of computational efficiency, we assume that we can define a finite set of candidate experimental designs before performing any model predictions. This is not a formal requirement, but it does allow us to avoid introducing an iterative optimization procedure to enable a computationally efficient implementation.

The remainder of this paper is organized as follows. In Section 2, we provide a high-level overview of the OED4P formulation based on push-forward measures. This section is intended to help the reader build intuition before we introduce the problem formally. In Section 3, we introduce our notation and formally define the forward and inverse problems considered in this work. In Section 4, we consider the special case of linear maps and Gaussian distributions. In Section 5, we define and discuss the OED4P problem, and Section 6 provides several numerical examples. Concluding remarks are in Section 7.

2. Motivation

We use a simple conceptual example to motivate the utility of optimal experimental design for prediction (OED4P) and demonstrate how two different experiments with similar reductions in initial estimates of uncertainty for parameters can have drastically different reductions in uncertainty for predictions. This simple example clearly illustrates how OED4P can be used to select the experiments that lead to the largest expected reduction in uncertainty for predictions.

OED4P requires both an *observational model* and *prediction model*. The observational model produces estimates of experimental data and the prediction model produces estimates of the QoI. These models can be used to obtain a set of outputs from the joint experiment-QoI data space by evaluating each model on a set of samples from an initial probability measure describing uncertainties in model input parameters. This output sample set

is referred to as the push-forward of the initial input sample set. An example of such a push-forward sample set is depicted in Figure 1 (top-left). In this example the observational and prediction model both produce scalar quantities.

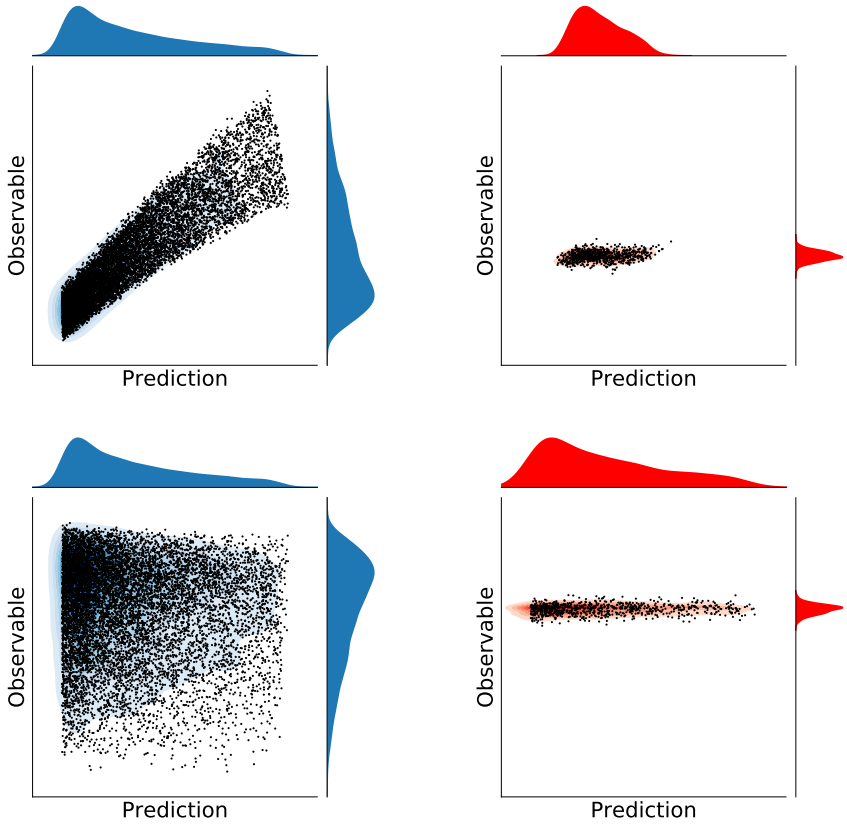

Figure 1: On the top row, the push-forward of the initial density (top-left) and the updated density (top-right) into the joint space defined by the observable and prediction data. On the bottom row, the push-forward of the initial density (bottom-left) and the updated density (bottom-right) into the joint space defined by a less-informative observable and prediction data.

Using data to reduce estimates of prediction (or parameter uncertainty) requires conditioning the aforementioned joint push-forward density on the available data. In the following, we assume that observational data are used to construct a probability measure that quantifies uncertainties on the

observable model output. Using this so-called *observed probability measure* it is possible to select a subset of samples from the push-forward sample set (e.g., using rejection sampling) such that that the observable components of this output subset are independent identically distributed (i.i.d.) samples from the observed probability measure. This sample set is said to be *data–consistent*.

The left plots of Figure 1 depict two observed joint observed-QoI densities, and their marginals, obtained using two different observational model outputs. The right plots of this figure depict the two updated joint densities, and marginals, after an observed measure for each model output is used to produce a data-consistent set of samples from the updated joint observed-QoI joint density. When the observed probability measure (red marginal on vertical axis) has less uncertainty than the push-forward of the initial measure (blue marginal on vertical axis) on the model input parameters (e.g., as measured by a reduction in variance), the predicted components of the data-consistent sample set may also exhibit reduced uncertainty. The amount of uncertainty reduced depends significantly on the relationship between the observable output and the prediction QoI. In the top row of Figure 1, we see a significant reduction in prediction uncertainty by either comparing the data-consistent (top-right) and push-forward (top-left) sample sets or their corresponding marginal densities shown on the top (for the prediction QoI) and right (for the observable output) of the axes. In contrast, there is almost no reduction in prediction uncertainty reduction using the observable output depicted in the bottom row of Figure 1. While the data–consistent sample set produces similar reductions in uncertainty in the observable component, there is almost no reduction in uncertainty for the prediction due to the weak correlation between this observable output and prediction. Thus, while such observational data may lead to significant reductions in input parameter uncertainty, we have failed to reduce the prediction uncertainty. Qualitatively speaking, the experiment used for the top row of Figure 1 provides a greater reduction in prediction uncertainty than the experiment used for the bottom row.

OED4P requires quantitative metrics to select the set of experiments which maximize the information gained from a set of experiments. While many metrics are reasonable, we use the Kullback-Leibler (KL) divergence [32] between initial and updated prediction densities. The KL divergence has several interpretations depending on the context in which it is utilized. For example, in an information theory context, the KL divergence is connected to quantities such as Shannon entropy [12]. In Bayesian contexts, the KL divergence quantifies the information gained in moving from a prior to pos-

terior distribution [19]. In the context of this work, the relationships to machine learning and coding theory are perhaps the most appropriate for interpreting how we utilize the KL divergence. Specifically, we use the KL divergence to quantify the information gained if one measure is used in place of another, e.g., if the observed measure replaces the predicted measure. This is also related to the efficiency in using samples from one measure to generate samples from a different measure, e.g., in using predicted samples and rejection sampling to generate samples from the observed measure.

We return to Figure 1 to provide additional context for the interpretation of the KL divergence. Specifically, using the KL divergence to measure the changes between the initial uninformed prediction-marginal, in the left plots of Figures 1, and the data-informed prediction-marginal of the same plots supports our qualitative conclusions. Here, the KL-divergence between the densities in the top row of Figure 1 is approximately 0.6532 which is significantly higher than the value of 0.0084 obtained for the densities in the bottom row, indicating that the first experiment is more informative.[2]

Since solutions to the OED or OED4P problems are sought prior to new or additional experimentation, the observational model is required to simulate the types of data that are likely to be observed. We therefore define an appropriate space of candidate distributions (e.g., Normal distributions centered around each potential datum) for each observable model output (the marginal distributions on the vertical axes of Figures 1 and then compute some meaningful statistic from this space of distributions. The most common approach is to simply take the average over this space, giving the expected information gained. In [33], the OED problem is solved by observations that maximize the expected information gained, measured by the KL divergence, between the initial (prior) and updated (posterior) densities on model input parameters. Due to the formulation of the stochastic inverse problem employed in [33], this is equivalent to maximizing the expected KL divergence between the push-forward of the initial density and the observed density, i.e., optimizing the expected difference between the distributions on the vertical axis in Figure 1. In this paper, we seek to maximize the expected information gained between the marginals associated with the push-forward of the initial density and the push-forward of the updated density in the space of predictions. In other words, we seek to optimize the expected difference between the distributions on the horizontal axis in Figure 1.

[2]The KL divergence is always non-negative and is minimized at zero when comparing identical distributions.

3. Data–informed prediction

In this section, we give the precise terminology and notation required to formally define the OED and OED4P problems, which are presented in Section 5. After introducing these preliminaries, we discuss the forward and inverse problems considered in this paper as well as a fundamental relationship between them that is exploited for the inference-for-prediction problem.

3.1. The spaces, maps, and some assumptions

Suppose the observational and prediction models have a common set of model input parameters denoted by $\lambda \in \Lambda \subset \mathbb{R}^p$. The set Λ represents the largest physically meaningful domain of parameter values. We assume that the set of prediction QoI are defined as functionals on the solution space of the prediction model. Furthermore, we assume these QoI, denoted by $\{Q_i(\lambda)\}_{i=1}^q$, are both fixed and known *a priori*. Then, we define the parameter-to-predictions map by $Q(\lambda) := (Q_1(\lambda), \cdots, Q_q(\lambda))^\top$. Let $\mathcal{Q} := Q(\Lambda) \subset \mathbb{R}^q$ denote the set of all possible predicted QoI data.

The goal is to choose a parameter-to-observables map for which we collect data to improve predictions on $\mathcal{Q}$. Assuming we can collect data on d observable outputs of the model, we let $\mathcal{E}$ represent the space of all potential d-dimensional parameter-to-observable maps for which we may collect data during experiments. The space $\mathcal{E}$ is referred to as the design space since it defines the space of all possible *experimental* designs. Let, $D \in \mathcal{E}$ be a specific design defined by the parameter-to-observable map $D(\lambda) := (D_1(\lambda), \cdots, D_d(\lambda))^\top$ for a particular set of observable output quantities, $\{D_j(\lambda)\}_{j=1}^d$. By $\mathcal{D} := D(\Lambda)$ we denote the range of the design D, which defines the set of all potentially observable data for this particular map. Note that the set of points defining $\mathcal{D}$ depends upon the map D, but as we will see, this dependence has no significant consequence in the analysis or algorithms, so we avoid further mention of it in this work. For problems where the elements in $\mathcal{E}$ are enumerated, we denote the kth design as $D^{(k)}$.

We assume $(\Lambda, \mathcal{B}_\Lambda, \mu_\Lambda)$, $(\mathcal{Q}, \mathcal{B}_\mathcal{Q}, \mu_\mathcal{Q})$, and (for each $D \in \mathcal{E}$) $(\mathcal{D}, \mathcal{B}_\mathcal{D}, \mu_\mathcal{D})$ are measure spaces with $\mathcal{B}_\Lambda$, $\mathcal{B}_\mathcal{Q}$, and $\mathcal{B}_\mathcal{D}$ denoting the Borel σ-algebras inherited from the metric topologies on $\Lambda \subset \mathbb{R}^p$, $\mathcal{Q} \subset \mathbb{R}^q$, and $\mathcal{D} \subset \mathbb{R}^d$, respectively. The measures μ_Λ, $\mu_\mathcal{Q}$, and $\mu_\mathcal{D}$ are the dominating measures for which probability densities (i.e., Radon-Nikodym derivatives of probability measures) are defined on each space. We also assume that the parameter-to-observables map, D, and the parameter-to-predictions map, Q, are at least piecewise smooth implying that D and Q are measurable maps.

3.2. Forward Problems

The forward uncertainty propagation problem considered in this work is elegantly described in terms of constructing a push-forward probability measure. First, let $\mathbb{P}_{\Lambda}^{\text{init}}$ denote an *initial* probability measure on $(\Lambda, \mathcal{B}_{\Lambda})$. The knowledgeable reader may think of this initial probability measure as a prior on the model input parameters, but we follow the nomenclature used in [10] to emphasize the fact that we ultimately solve a different inverse problem from the classical Bayesian formulation with the aid of this measure. Since D is measurable, it induces a push-forward measure on $\mathbb{P}_{\mathcal{D}}^{D(\text{init})}$ on $\mathcal{D}$.

Definition 1 (The Initial Push-Forward Measure for Observables). *Given an initial probability measure, $\mathbb{P}_{\Lambda}^{init}$, on $(\Lambda, \mathcal{B}_{\Lambda})$ that is absolutely continuous with respect to μ_{Λ} and admits a density π_{Λ}^{init}, the forward problem for observables seeks to determine the push-forward probability measure $\mathbb{P}_{\mathcal{D}}^{D(init)}$ on $(\mathcal{D}, \mathcal{B}_{\mathcal{D}})$ such that for all $B \in \mathcal{B}_{\mathcal{D}}$,*

$$\mathbb{P}_{\mathcal{D}}^{D(init)}(B) = \mathbb{P}_{\Lambda}^{init}(D^{-1}(B)) = \int_{D^{-1}(B)} \pi_{\Lambda}^{init} \, d\mu_{\Lambda}. \tag{1}$$

This push-forward probability measure on observables is a key component of the formulation for solving the stochastic inverse problem described in the next section. In a similar manner, the fact that Q is measurable implies that the initial probability measure and the map, Q, induces a push-forward measure $\mathbb{P}_{\mathcal{Q}}^{Q(\text{init})}$ on $\mathcal{Q}$.

Definition 2 (The Initial Push-Forward Measure for Predictions). *Given an initial probability measure, $\mathbb{P}_{\Lambda}^{init}$, on $(\Lambda, \mathcal{B}_{\Lambda})$ that is absolutely continuous with respect to μ_{Λ} and admits a density π_{Λ}^{init}, the forward problem for predictions seeks to determine the push-forward probability measure $\mathbb{P}_{\mathcal{Q}}^{Q(init)}$ on $(\mathcal{Q}, \mathcal{B}_{\mathcal{Q}})$ such that for all $A \in \mathcal{B}_{\mathcal{Q}}$,*

$$\mathbb{P}_{\mathcal{Q}}^{Q(init)}(A) = \mathbb{P}_{\Lambda}^{init}(Q^{-1}(A)) = \int_{Q^{-1}(A)} \pi_{\Lambda}^{init} \, d\mu_{\Lambda}. \tag{2}$$

We additionally assume that $\mathbb{P}_{\mathcal{D}}^{D(\text{init})}$ and $\mathbb{P}_{\mathcal{Q}}^{Q(\text{init})}$ are absolutely continuous with respect to $\mu_{\mathcal{D}}$ and $\mu_{\mathcal{Q}}$ and admit densities $\pi_{\mathcal{D}}^{D(\text{init})}$ and $\pi_{\mathcal{Q}}^{Q(\text{init})}$, respectively. In the context of the discussion in Section 2, $\pi_{\mathcal{D}}^{D(\text{init})}$ is illustrated by the marginal densities shown on the right of the vertical axes and $\pi_{\mathcal{Q}}^{Q(\text{init})}$ is illustrated by the marginal densities shown on the top of the horizontal axes in Figure 1.

3.3. A Stochastic Inverse Problem

The formulation of a stochastic inverse problem which utilizes initial (prior) estimates of uncertainty on data is essential for understanding both the data-informed prediction and OED4P problems we define. In this section, we summarize this inverse problem.

At a conceptual level, assume we are given a distribution on observational data (e.g., obtained from experimentation) and an initial distribution on the input parameters. Then, the inverse problem we consider seeks an updated measure on parameters which is data-consistent in the sense that its subsequent push-forward back through the parameter-to-observables map matches the observed measure. In other words, the updated measure on parameters is a *pullback* of the observed measure. This is defined more precisely below.

Definition 3 (Stochastic Inverse Problem). *Given a probability measure $\mathbb{P}_{\mathcal{D}}^{obs}$ on $(\mathcal{D}, \mathcal{B}_{\mathcal{D}})$ that is absolutely continuous with respect $\mu_{\mathcal{D}}$ and admits a density $\pi_{\mathcal{D}}^{obs}$, the inverse problem seeks to determine a probability measure $\mathbb{P}_{\Lambda}$ on $(\Lambda, \mathcal{B}_{\Lambda})$ that is absolutely continuous with respect to μ_{Λ} and admits a probability density π_{Λ}, such that the subsequent push-forward measure induced by the map, D, is data-consistent in the sense that*

$$\mathbb{P}_{\Lambda}(D^{-1}(A)) = \int_{D^{-1}(A)} \pi_{\Lambda}\, d\mu_{\Lambda} = \mathbb{P}_{\mathcal{D}}^{D}(A) = \mathbb{P}_{\mathcal{D}}^{obs}(A) = \int_{A} \pi_{\mathcal{D}}^{obs}\, d\mu_{\mathcal{D}}, \qquad (3)$$

for any $A \in \mathcal{B}_{\mathcal{D}}$.

There may be multiple probability measures that are data-consistent[3] in the sense of Definition 3. This is analogous to an under-determined deterministic inverse problem where multiple parameters may produce the same observed datum. A unique solution may be obtained by imposing additional constraints or structure on the stochastic inverse problem. In this work, we follow [9] and impose such structure by specifying an initial probability measure of the model parameters. Specifically, we introduce an *initial* probability measure $\mathbb{P}_{\Lambda}^{\text{init}}$ on $(\Lambda, \mathcal{B}_{\Lambda})$ that is absolutely continuous with

[3]The notion of data-consistency described in Definition 3 is fundamentally different from the traditional notion of consistency is Bayesian analysis. See Remark 1 for futher details.

respect to μ_Λ and admits a probability density $\pi_\Lambda^{\text{init}}$.[4] Then, to guarantee the existence and uniqueness of a solution to the inverse problem, given in Definition 3, we make the following assumption.

Assumption 1. *There exists a constant $C > 0$ such that $\pi_{\mathcal{D}}^{obs}(y) \leq C\pi_{\mathcal{D}}^{D(init)}(y)$ for a.e. $y \in \mathcal{D}$.*

Since the observed density and the model are assumed to be fixed, this is only an assumption on the initial measure. We sometimes refer to this assumption as the *Predictability Assumption* since it implies that any output event with non-zero *observed* probability has a non-zero *predicted probability* defined by the push-forward of the initial measure. This is actually the same constant that one would estimate if using rejection sampling to produce samples from $\pi_{\mathcal{D}}^{\text{obs}}$ using samples generated from $\pi_{\mathcal{D}}^{D(\text{init})}$ (see [9] for details).

Under the predictability assumption, the unique solution to the stochastic inverse problem is obtained using a disintegration of measures [14] along with Bayes' rule. This solution is referred to as the *updated* probability measure, denoted $\mathbb{P}_\Lambda^{\text{up}}$, and is given by

$$\mathbb{P}_\Lambda^{\text{up}}(A) = \int_{\mathcal{D}} \left(\int_{A \cap D^{-1}(y)} \pi_\Lambda^{\text{init}}(\lambda) \frac{\pi_{\mathcal{D}}^{\text{obs}}(D(\lambda))}{\pi_{\mathcal{D}}^{D(\text{init})}(D(\lambda))} d\mu_{\Lambda,y}(\lambda) \right) d\mu_{\mathcal{D}}(y), \ \forall A \in \mathcal{B}_\Lambda. \tag{4}$$

In terms of probability density functions, the updated density is defined as

$$\pi_{\mathbf{\Lambda}}^{\text{up}}(\lambda) := \pi_{\mathbf{\Lambda}}^{\text{init}}(\lambda) \frac{\pi_{\mathcal{D}}^{\text{obs}}(Q(\lambda))}{\pi_{\mathcal{D}}^{Q(\text{init})}(Q(\lambda))}, \quad \lambda \in \mathbf{\Lambda}. \tag{5}$$

The solution to the stochastic inverse problem defined in Definition 3 given in (5) is asymptoptically consistent. In [9], we prove that the error in the updated density, measured in the L^1 or total variation norm, is bounded by the error in the observed density. Consequently the updated density converges to a "true" updated density in the limit of increasing data as the characterization of $\pi_{\mathcal{D}}^{\text{obs}}$ converges. For a more detailed derivation, including discussions on stability in the total variation metric, we refer the interested reader to [9].

[4] In [9], the initial and updated measures were referred to as the prior and probability measures. In this paper we follow the exposition in [10] and refer to the prior and posterior as the initial and updated measures/densities, respectively, to reinforce the conceptual differences of the inverse problem used in this paper and classical Bayesian inference.

Remark 1. *The aforementioned notion of consistency for the stochastic inverse problem in Definition 3 is fundamentally different from the traditional notion of consistency is Bayesian analysis. Specifically, the consistency of the former is defined in terms of a pullback probability measure rather than an asymptotic convergence towards the "true" parameter value in the limit of increasing data.*

3.4. The Data-Informed Prediction Problem

Given a distribution on the observations, a data-informed prediction conceptually follows a two step procedure: (i) solve the stochastic inverse problem of Definition 3 by computation of the updated density given in (5); and (ii) use the parameter-to-predictions map to compute the push-forward of the updated measure for the QoI. We define this push-forward formally as the data-informed prediction.

Definition 4 (Data-Informed Prediction). *Given an updated probability measure $\mathbb{P}_{\Lambda}^{up}$ on $(\Lambda, \mathcal{B}_{\Lambda})$ that is absolutely continuous with respect to μ_{Λ} and admits a density π_{Λ}^{up}, the data-informed prediction seeks the push-forward probability measure $\mathbb{P}_{\mathcal{Q}}^{Q(up)}$ on $(\mathcal{Q}, \mathcal{B}_{\mathcal{Q}})$ such that for all $A \in \mathcal{B}_{\mathcal{Q}}$,*

$$\mathbb{P}_{\mathcal{Q}}^{Q(up)}(A) = \mathbb{P}_{\Lambda}^{up}(Q^{-1}(A)). \tag{6}$$

As before, we assume $\mathbb{P}_{\mathcal{Q}}^{Q(\mathrm{up})}$ is absolutely continuous with respect to $\mu_{\mathcal{Q}}$ and admits a density $\pi_{\mathcal{Q}}^{Q(\mathrm{up})}$.

Given an i.i.d. sample set of parameters from the initial density, we can algorithmically follow the two step procedure as follows. First, use the parameter-to-observations map to approximate the push-forward of the initial density on the observation space. Then, use the ratio of the observed density to this push-forward to perform rejection sampling in the observation space. Keeping track of indices of samples that are accepted in the observation space, a subset of the initial sample set of parameters can be identified as a set of i.i.d. samples from the updated density on parameter space. This solves the first step of obtaining a solution to the stochastic inverse problem, which is described as its own algorithm that is analyzed in [9]. The second step is then solved by propagating these i.i.d. samples from the updated density through the parameter-to-predictions map to obtain a set of i.i.d. samples from $\pi_{\mathcal{Q}}^{Q(\mathrm{up})}$.

Algorithm 1 gives an alternative approach to obtaining i.i.d. samples from $\pi_{\mathcal{Q}}^{Q(\mathrm{up})}$ that bypasses the process of obtaining a solution to the stochastic inverse problem. A necessary input to this algorithm is a set of *joint* samples in the observation and prediction spaces coming from an initial sample

set of parameters. In other words, the parameter-to-predictions map must have already been applied to the initial sample set. Assuming this is the case, then all the computations occur only in the observed data space, so it is not technically necessary to know the actual sample set of parameters from the input distribution. This implies that the Data-Informed Prediction problem is solvable even when the initial parameter distribution is unknown.

Algorithm 1: Generating Samples from the Data-Informed Prediction

Input:

1. $\{(D(\lambda^{(i)}), Q(\lambda^{(i)}))\}_{i=1}^{N}$ where $\{\lambda^{(i)}\}_{i=1}^{N} \sim \pi_{\Lambda}^{\text{init}}$.
2. $\pi_{\mathcal{D}}^{\text{obs}}$.

Pre-processing computations::

1. Estimate $\pi_{\mathcal{D}}^{D(\text{init})}$ and define $r(\lambda) := \pi_{\mathcal{D}}^{\text{obs}}(D(\lambda))/\pi_{\mathcal{D}}^{D(\text{init})}(D(\lambda))$.
2. Estimate $M := \max_{\Lambda} r(\lambda) \approx \max_{1 \le i \le N} r(\lambda^{(i)})$.

for $i = 1, \ldots, N$ **do**
  Generate a random number, u, from a uniform distribution on $[0, 1]$;
  Compute the ratio: $\eta = r(D(\lambda^{(i)}))/M$;
  if $\eta > u$ **then**
    Accept $Q(\lambda^{(i)})$;
  else
    Reject $Q(\lambda^{(i)})$;
  end
end

Output: Accepted QoI samples.

While this algorithm is computationally convenient, it obfuscates how the updated measure forms the underlying connection between the data-informed measure on predictions and the observed measure. From Definition 4, the connection between the data-informed measure on predictions and the observed measure is clear and allows us draw two useful general inferences on data-informed probabilities. First, for any $A \in \mathcal{B}_{\mathcal{Q}}$, we know that $Q^{-1}(A) \in \mathcal{B}_{\Lambda}$ and that $Q^{-1}(A) \subseteq D^{-1}(D(Q^{-1}(A)))$, which implies

$$\mathbb{P}_{\mathcal{Q}}^{Q(\text{up})}(A) = \mathbb{P}_{\Lambda}^{\text{up}}(Q^{-1}(A)) \le \mathbb{P}_{\Lambda}^{\text{up}}(D^{-1}(D(Q^{-1}(A)))) = \mathbb{P}_{\mathcal{D}}^{\text{obs}}(D(Q^{-1}(A))).$$

This inequality describes how the observations bound probabilities of prediction events from above. By reversing the roles of Q and D, we also conclude

that for any $A \in \mathcal{B}_{\mathcal{D}}$,

$$\mathbb{P}_{\mathcal{D}}^{\text{obs}}(A) \leq \mathbb{P}_{\mathcal{Q}}^{Q(\text{up})}(Q(D^{-1}(A)).$$

This inequality describes how certain prediction events are bounded below in probability.

4. A Special Case: Linear Gaussian Models

The purpose of this section is to provide some additional intuition about how the structures of the observation and prediction measures are related in the special case where all maps are linear and the initial and observed densities are Gaussian. Thus, in this section we assume that $Q : \Lambda \to \mathcal{Q}$ and $D : \Lambda \to \mathcal{D}$ are linear maps, i.e., $Q \in \mathbb{R}^{q \times p}$ and $D \in \mathbb{R}^{d \times p}$, and that $\pi_{\Lambda}^{\text{init}} \sim N(0, \Sigma_{\Lambda})$ and $\pi_{\mathcal{D}}^{\text{obs}} \sim N(0, \Sigma_{\mathcal{D}})$.

Assuming further that $p \geq \max\{q, d\}$ (i.e., there are at least as many parameters as either observables of QoI) and that the matrices defining Q and D both have full rank, the updated density as well as the push-forwards of the initial and updated densities through either map are also given by normal distributions with the following covariance matrices:

$$\begin{aligned}
\text{cov}\left(\pi_{\mathcal{D}}^{D(\text{init})}\right) &= \Sigma_{\mathcal{D}}^{D(\text{init})} = D\Sigma_{\Lambda}D^{\top} \\
\text{cov}\left(\pi_{\mathcal{Q}}^{Q(\text{init})}\right) &= \Sigma_{\mathcal{Q}}^{Q(\text{init})} = Q\Sigma_{\Lambda}Q^{\top} \\
\text{cov}\left(\pi_{\Lambda}^{\text{up}}\right) &= \Sigma_{\Lambda}^{\text{up}} = \left(\Sigma_{\Lambda}^{-1} + D^{\top}\left[\Sigma_{\mathcal{D}}^{-1} - \left(D\Sigma_{\Lambda}D^{\top}\right)^{-1}\right]D\right)^{-1} \\
\text{cov}\left(\pi_{\mathcal{D}}^{D(\text{up})}\right) &= \Sigma_{\mathcal{D}}^{D(\text{up})} = D\Sigma_{\Lambda}^{\text{up}}D^{\top} \\
\text{cov}\left(\pi_{\mathcal{Q}}^{Q(\text{up})}\right) &= \Sigma_{\mathcal{Q}}^{Q(\text{up})} = Q\Sigma_{\Lambda}^{\text{up}}Q^{\top}.
\end{aligned}$$

From these expression we can make two additional observations to provide further insight into the structure of the updated solution:

- If $D\Sigma_{\Lambda}D^{\top} = \Sigma_{\mathcal{D}}$, then $\Sigma_{\Lambda}^{\text{up}} = \Sigma_{\Lambda}$. In other words, if the push-forward of the initial density already matches the observed density, then the updated density is identical to the initial density.
- If D is invertible, the stochastic inverse problem has a unique solution with a covariance given by $\Sigma_{\Lambda}^{\text{up}} = \left(D^{\top}\Sigma_{\mathcal{D}}^{-1}D\right)^{-1}$ since $D^{\top}(D\Sigma_{\Lambda}D^{\top})^{-1}D = \Sigma_{\Lambda}^{-1}$. In other words, the updated covariance does not depend on the initial density.

It is convenient to use the Woodbury identity to rewrite the updated covariance as

$$\left(\Sigma_\Lambda^{-1} + D^\top \left[\Sigma_\mathcal{D}^{-1} - \left(D\Sigma_\Lambda D^\top\right)^{-1}\right] D\right)^{-1} = \Sigma_\Lambda - \Sigma_\Lambda D^\top M^{-1} D\Sigma_\Lambda$$

where

$$M = \Sigma_\mathcal{D} + D\Sigma_\Lambda D^\top - \Sigma_\mathcal{D}\left(I - D\Sigma_\Lambda D^\top \Sigma_\mathcal{D}^{-1}\right)^{-1},$$

so we can write

$$\Sigma_\Lambda^{\text{up}} = \Sigma_\Lambda - \Sigma_\Lambda D^\top M^{-1} D\Sigma_\Lambda \tag{7}$$

$$\Sigma_\mathcal{Q}^{Q(\text{up})} = Q\Sigma_\Lambda Q^\top - Q\Sigma_\Lambda D^\top M^{-1} D\Sigma_\Lambda Q^\top. \tag{8}$$

This form exposes two important quantities: (i) the operator $\Sigma_\Lambda D^\top : \mathcal{D} \to \mathcal{Q}$ that denotes the initial-weighted observable-to-parameter map and determines the perturbation of π_Λ^{up} from $\pi_\Lambda^{\text{init}}$; and (ii) $Q\Sigma_\Lambda D^\top : \mathcal{D} \to \mathcal{Q}$ that denotes the initial-weighted observable-to-prediction map and determines the perturbation of $\pi_\mathcal{Q}^{Q(\text{up})}$ from $\pi_\mathcal{Q}^{Q(\text{init})}$.

It is well-known that the covariance of the posterior computed using the classical Bayesian approach is given by $\left(\Sigma_\Lambda^{-1} + D^\top \Sigma_\mathcal{D}^{-1} D\right)^{-1}$ which leads to a very similar expression for the push-forwards of the posterior, but with $M = \Sigma_\mathcal{D} + D\Sigma_\Lambda D^\top$. In both cases, the initial-weighted observable-to-prediction map strongly influences the information gained from the push-forward of the initial density, $\pi_\mathcal{Q}^{Q(\text{init})}$, to the push-forward of the updated density, $\pi_\mathcal{Q}^{Q(\text{up})}$.

5. Optimal Experimental Design

In this section, we define the OED and OED4P problems based upon the stochastic inverse problem and data-based prediction problems defined in Definitions 3 and 4, respectively. We first summarize the concept of information gain, using the KL divergence between two probability densities, which is necessary for ranking the quality of different designs. Then, we define both the OED and OED4P problems in terms of optimizing the expected information gain over a specific space of possible observed densities.

5.1. Information gain: Kullback-Leibler divergence

To quantify the *information gain* of a design, we use the Kullback-Leibler (KL) divergence [32]. As mentioned in Section 2, the interpretation of the KL divergence depends greatly upon the context in which it is applied.

In the context of this work, it quantifies the information gain achieved if one probability measure (e.g., the updated probability measure) replaces another (e.g., the initial probability measure). Moreover, it has a practical interpretation in terms of the efficiency in using samples from one probability measure (e.g., the initial probability measure) to generate samples from another probability measure (e.g., the updated probability measure).

For the sake of a general discussion, let π_1 and π_2 be two probability densities on a measure space $(\mathbb{X}, \mathcal{B}, \mu)$. The KL divergence from π_1 to π_2 is formally defined as

$$\mathrm{KL}(\pi_2 \| \pi_1) := \int_{\mathbb{X}} \pi_2 \log\left(\frac{\pi_2}{\pi_1}\right) d\mu. \tag{9}$$

In general, it may be computationally infeasible to accurately approximate the integral in Eq. (9) especially if $\mathbb{X}$ is high-dimensional. However, if the measure associated with π_2, denoted by $\mathbb{P}_2$, is absolutely continuous with respect to the measure associated with π_1, denoted by $\mathbb{P}_1$, then by standard results for Radon–Nikodym derivatives, we have that μ–a.e.,

$$\pi_2 = \frac{d\mathbb{P}_2}{d\mu} = \frac{d\mathbb{P}_2}{d\mathbb{P}_1}\frac{d\mathbb{P}_1}{d\mu} = \frac{d\mathbb{P}_2}{d\mathbb{P}_1}\pi_1.$$

Let the Radon–Nikodym derivative of $\mathbb{P}_2$ with respect to $\mathbb{P}_1$ be denoted by

$$r := \frac{d\mathbb{P}_2}{d\mathbb{P}_1},$$

which defines a non–negative measurable function in $L^1(\mathbb{X})$, then, we can rewrite (9) as

$$\mathrm{KL}(\pi_2 \| \pi_1) := \int_{\mathbb{X}} r \log(r)\, d\mathbb{P}_1. \tag{10}$$

Estimating this integral then becomes tractable with Monte Carlo sampling assuming it is straightforward to generate i.i.d. samples following the $\mathbb{P}_1$ distribution.

In this paper, we are primarily interested in two applications of the KL divergence between densities. The first is the KL divergence between the initial density, $\pi_\Lambda^{\mathrm{init}}$, and the updated density, $\pi_\Lambda^{\mathrm{up}}$, given by

$$\mathrm{KL}(\pi_\Lambda^{\mathrm{up}} \| \pi_\Lambda^{\mathrm{init}}) = \int_\Lambda \pi_\Lambda^{\mathrm{up}}(\lambda) \log\left(\frac{\pi_\Lambda^{\mathrm{up}}(\lambda)}{\pi_\Lambda^{\mathrm{init}}(\lambda)}\right) d\mu_\Lambda = \int_\Lambda r(\lambda) \log(r(\lambda))\ d\mathbb{P}_\Lambda^{\mathrm{init}},$$

where $r(\lambda)$ is given by

$$r(\lambda) = \frac{\pi_{\mathcal{D}}^{\text{obs}}(D(\lambda))}{\pi_{\mathcal{D}}^{D(\text{init})}(D(\lambda))}.$$

In our approach for solving the stochastic inverse problem, samples are generated from the initial density and the corresponding model evaluations are used to estimate $\pi_{\mathcal{D}}^{D(\text{init})}$ and, in turn, $r(D(\lambda))$. Thus, we already have the ability to approximate $\text{KL}(\pi_{\Lambda}^{\text{up}} \| \pi_{\Lambda}^{\text{init}})$ by integrating

$$r(\lambda) \log(r(\lambda))$$

with respect to the initial probability measure.

The second application is the KL divergence between the push-forward of the initial density for predictions, $\pi_{\mathcal{Q}}^{Q(\text{init})}$, and the push-forward of the updated density for prediction, $\pi_{\mathcal{Q}}^{Q(\text{up})}$, which is given by

$$\begin{aligned}\text{KL}(\pi_{\mathcal{Q}}^{Q(\text{up})} \| \pi_{\mathcal{Q}}^{Q(\text{init})}) &= \int_{\mathcal{Q}} \pi_{\mathcal{Q}}^{Q(\text{up})}(q) \log\left(\frac{\pi_{\mathcal{Q}}^{Q(\text{up})}(q)}{\pi_{\mathcal{Q}}^{Q(\text{init})}(q)}\right) d\mu_{\mathcal{Q}} \\ &= \int_{\mathcal{Q}} r_{\mathcal{Q}}(q) \log\left(r_{\mathcal{Q}}(q)\right) \; d\mathbb{P}_{\mathcal{Q}}^{Q(\text{init})},\end{aligned}$$

where $r_{\mathcal{Q}}(q)$ is defined as

$$r_{\mathcal{Q}}(q) = \frac{\pi_{\mathcal{Q}}^{Q(\text{up})}(q)}{\pi_{\mathcal{Q}}^{Q(\text{init})}(q)}.$$

Again, since we have already generated samples from the initial density and computed the corresponding model evaluations, we have samples from $\pi_{\mathcal{Q}}^{Q(\text{init})}$ and can approximate the integral of

$$r_{\mathcal{Q}}(q) \log\left(r_{\mathcal{Q}}(q)\right)$$

with respect to $\mathbb{P}_{\mathcal{Q}}^{Q(\text{init})}$.

5.2. Expected information gain

Ideally, the solution to the OED problem determines a design before experimental data are collected. Yet, computation of either $\text{KL}(\pi_{\Lambda}^{\text{up}} \| \pi_{\Lambda}^{\text{init}})$ or $\text{KL}(\pi_{\mathcal{Q}}^{Q(\text{up})} \| \pi_{\mathcal{Q}}^{Q(\text{init})})$ requires specification of an observed density. Letting $\mathcal{C}$ denote the space of all potential observed densities, the goal is to define the

expected information gain as some kind of average over $\mathcal{C}$ in a meaningful way. However, this is far too general of a space to use to define the expected information gain. This space includes densities that are unlikely to be observed in reality. Therefore, we restrict $\mathcal{C}$ to be a space more representative of densities that may be observed in reality in such a way that we can also prescribe a relative likelihood on these densities. In other words, we define $\mathcal{C}$ such that a probability measure/density can be prescribed on $\mathcal{C}$ which will allow us to compute an expectation.

With no experimental data available to specify an observed density, we assume for simplicity that the observed density will belong to a parametric family of distributions parameterized by a finite-dimensional vector. We denote such a parametric family by $\mathcal{P}(\boldsymbol{h})$ where the parameters used to define the distributions are denoted by $\boldsymbol{h} \in \mathcal{H} \subset \mathbb{R}^h$. We refer to these h–dimensional parameter vectors $\boldsymbol{h}$ as hyperparameters to distinguish them from the model input parameters denoted by λ, and $\mathcal{H}$ denotes the space of all plausible values considered for these hyperparameters. A simple example of such a parametric family is the family of Gaussian distributions on $\mathcal{D}$ with hyperparameters given by the d–dimensional mean and a diagonal $d \times d$ covariance matrix so that the dimension of the hyperparameters, h, is at most $2d$. With this general notation, we define

$$\mathcal{C} := \left\{ \pi_{\mathcal{D}}^{\text{obs}}(\boldsymbol{d}; \boldsymbol{h}) \in \mathcal{P}(\boldsymbol{h}) : \boldsymbol{d} \in \mathcal{D}, \boldsymbol{h} \in \mathcal{H} \right\}, \tag{11}$$

where we write $\pi_{\mathcal{D}}^{\text{obs}}(\boldsymbol{d}; \boldsymbol{h})$ to make the dependence of the observed density on the specification of the hyperparameter vector $\boldsymbol{h}$ explicit.

Note that with this definition, the elements of $\mathcal{C}$, defined as functions of $\boldsymbol{d}$, are in 1–to–1 correspondence with $\mathcal{H}$. Let $\beta : \mathcal{C} \to \mathcal{H}$ denote the bijection defined by $\beta(\pi_{\mathcal{D}}^{\text{obs}}(\boldsymbol{d}; \boldsymbol{h})) = \boldsymbol{h}$. This is a measurable function which induces a σ–algebra $\mathcal{B}_{\mathcal{C}}$ on $\mathcal{C}$ that is in 1–to–1 correspondence with the Borel σ–algebra $\mathcal{B}_{\mathcal{H}}$ naturally defined by the restriction of the h–dimensional Borel σ–algebra to $\mathcal{H}$. We can therefore use any probability measure or density defined on $(\mathcal{H}, \mathcal{B}_{\mathcal{H}})$ to describe probabilities or relative likelihoods on $(\mathcal{C}, \mathcal{B}_{\mathcal{C}})$. In other words, the computations of expected information gain can take place on the h–dimensional measurable space $(\mathcal{H}, \mathcal{B}_{\mathcal{H}})$ in place of $(\mathcal{C}, \mathcal{B}_{\mathcal{C}})$. We let $\mathbb{P}_{\mathcal{H}}$ denote a given probability measure on $(\mathcal{H}, \mathcal{B}_{\mathcal{H}})$ from which we can generate samples and compute an expected value. We then define the *expected information gain* (EIG) for the parameters, denoted $\mathbb{E}_{\mathcal{C}}\left(\text{KL}(\pi_{\Lambda}^{\text{up}} \| \pi_{\Lambda}^{\text{init}})\right)$, as just

described,

$$\mathbb{E}_{\mathcal{C}}\left(\mathrm{KL}(\pi_{\Lambda}^{\mathrm{up}}\|\pi_{\Lambda}^{\mathrm{init}})\right) := \int_{\mathcal{H}}\int_{\Lambda}\pi_{\Lambda}^{\mathrm{up}}(\lambda;\boldsymbol{h})\log\left(\frac{\pi_{\Lambda}^{\mathrm{up}}(\lambda;\boldsymbol{h})}{\pi_{\Lambda}^{\mathrm{init}}(\lambda)}\right)d\mu_{\Lambda}d\mathbb{P}_{\mathcal{H}}$$
$$= \int_{\mathcal{H}}\int_{\Lambda} r(\lambda;\boldsymbol{h})\log\left(r(\lambda;\boldsymbol{h})\right)d\mathbb{P}_{\Lambda}^{\mathrm{init}}d\mathbb{P}_{\mathcal{H}}, \quad (12)$$

where we make explicit that $\pi_{\Lambda}^{\mathrm{up}}$ and r are functions of the observed density and, by our restriction of the space of observed densities in Eq. (11), functions of $\boldsymbol{h} \in \mathcal{H}$.

Similarly, we define the *expected information gain for predictions* (EIG4P), denoted $\mathbb{E}_{\mathcal{C}}\left(\mathrm{KL}(\pi_{\mathcal{Q}}^{Q(\mathrm{up})}\|\pi_{\mathcal{Q}}^{Q(\mathrm{init})})\right)$, as

$$\mathbb{E}_{\mathcal{C}}\left(\mathrm{KL}(\pi_{\mathcal{Q}}^{Q(\mathrm{up})}\|\pi_{\mathcal{Q}}^{Q(\mathrm{init})})\right) := \int_{\mathcal{H}}\int_{\mathcal{Q}}\pi_{\mathcal{Q}}^{Q(\mathrm{up})}(q)\log\left(\frac{\pi_{\mathcal{Q}}^{Q(\mathrm{up})}(q)}{\pi_{\mathcal{Q}}^{Q(\mathrm{init})}(q;\boldsymbol{h})}\right)d\mu_{\mathcal{Q}}d\mathbb{P}_{\mathcal{H}}$$
$$= \int_{\mathcal{H}}\int_{\mathcal{Q}} r_{\mathcal{Q}}(q;\boldsymbol{h})\log\left(r_{\mathcal{Q}}(q;\boldsymbol{h})\right)d\mathbb{P}_{\mathcal{Q}}^{Q(\mathrm{init})}d\mathbb{P}_{\mathcal{H}} \quad (13)$$

where we make explicit that $\pi_{\mathcal{Q}}^{Q(\mathrm{up})}$ are $r_{\mathcal{Q}}$ are functions of the observed density and therefore functions of $\boldsymbol{h} \in \mathcal{H}$. In the numerical examples of Section 6, we describe the specific forms of $\mathcal{P}(\boldsymbol{h})$ that are assumed along with the probability measure $\mathbb{P}_{\mathcal{H}}$ used in each computation.

Computing either of these EIGs appears to be a computationally expensive procedure since it requires solving a large number of stochastic inverse problems and approximating $\pi_{\mathcal{D}}^{D(\mathrm{init})}$ can be expensive if we use a non-parametric kernel density estimation technique. However, our approach for solving stochastic inverse problems only requires approximating the push-forward of the *fixed* initial density. In other words, the push–forward is itself fixed across all choices of observed densities from $\mathcal{C}$. This implies that this single push-forward can be used to approximate updated densities for different observed densities without requiring additional model evaluations. We comment further on the scalability of this OED formulation in Section 5.4 where we outline the algorithmic procedures for solving the problems discussed in this work.

Remark 2. *The choice of the space of possible observed densities $\mathcal{C}$ can influence the designs generated by our OED framework. OED attempts to select informative data before the data is collected. Consequently, OED requires a statistical model for generating likely data. In this paper we do this*

via the specification of $\mathcal{C}$. The need for a generative data model is also made in Bayesian OED. These methods a priori specify the relationship between the model and data, e.g. the error is Gaussian with a certain mean and standard deviation.

5.3. Defining the OED and the OED4P

Recall that each experimental design is defined as a d-dimensional parameter-to-observables map computed from the model, and we seek the optimal map corresponding to a set of d measurement devices to deploy in the field. Given a physics-based model, initial information on the model parameters, a space of potential experimental designs, and a generic description of the uncertainties for each observable, we define the OED for information gained on parameters as

$$D_{\text{OED}} := \arg\max_{D\in\mathcal{E}} \left(\mathbb{E}_{\mathcal{D}} \left(\text{KL}(\pi_{\Lambda}^{\text{up}} \| \pi_{\Lambda}^{\text{init}}; D) \right) \right), \tag{14}$$

where we have explicitly denoted that the KL-divergence, and thus the EIG, depends on the design, D, chosen from the design space $\mathcal{E}$.

Similarly, we define the OED4P for information gained on predictions as

$$D_{\text{OED4P}} := \arg\max_{D\in\mathcal{E}} \left(\mathbb{E}_{\mathcal{D}} \left(\text{KL}(\pi_{\mathcal{Q}}^{Q(\text{up})} \| \pi_{\mathcal{Q}}^{Q(\text{init})}; D) \right) \right), \tag{15}$$

where, as above, we have explicitly denoted that the KL-divergence, and thus the EIG4P, depends on the design, D, chosen from the design space $\mathcal{E}$.

Following [33], we assume the design space consists of a finite number of candidate designs, and we evaluate the EIG and EIG4P for all of these to determine the OED and the OED4P, respectively. This assumption avoids issues associated with continuous design spaces, but it does have certain implications. In Section 5.4, we discuss how this provides some notion of scalability if this design space is determined a priori, but it also limits the number of experiments we can consider.

5.4. Scalability of the OED and OED4P

In general, OED is extremely computationally demanding as it requires solving numerous inverse problems to compute the EIG, and this is wrapped within an optimization routine to maximize the EIG. Scalable approaches for other OED formulations have been developed for linear inverse problems [2, 3] and for the classical Bayesian OED for prediction formulations [5].

Under certain conditions, the approach developed in [33] for OED and the approach developed in this paper for OED4P are also scalable. The dominant cost in solving the stochastic inverse problem employed here is the estimation of the push-forward of the initial density, $\pi_{\mathcal{D}}^{D(\text{init})}$. However, given an approximation of this push-forward, we can explore updated densities associated with different observed densities for minimal computational cost. Thus, for a given design, computing the EIG and the EIG4P cost approximately the same as solving a single stochastic inverse problem, which costs as much as solving a forward UQ problem. In addition, if the set of candidate designs and the predictions are known a priori, then the computational model only needs to be evaluated once for each sample from the initial density. In this scenario, determining the OED and the OED4P cost about the same as solving a forward UQ problem.

On the other hand, if a candidate design may contain multiple measurements, then the cardinality of the space of candidate designs can grow factorially. Thus, the approaches in [33] and in this paper for OED and OED4P are not scalable with the number of measurements. In addition, if the space of candidate observed densities contains distributions that are more difficult to invert than others, then the accuracy of the proposed approaches for OED and OED4P may not scale well as the costs in computing the EIG and EIG4P are then dictated by the most difficult inverse problems to solve to sufficient accuracy. In these cases, it may be possible to exploit the connections with deterministic optimization (as in [25]) to develop scalable approaches based on a continuous design space, but this is beyond the scope of this paper.

6. Numerical Results

In this section, we present some numerical results to illustrate the difference between OED and OED4P. The first example focuses on the difference between EIG and EIG4P for the case of linear parameter-to-observable and parameter-to-prediction maps with Gaussian initial and observed densities. The next three examples consider discretized (partial or ordinary) differential equations. For each of these three examples, both the parameter-to-observable map and the parameter-to-prediction map are nonlinear and we define the space of candidate observed densities to be an appropriately parameterized family of Gaussian distributions.

6.1. Inference-for-prediction for a Linear Gaussian Problem

In this section, we consider a linear map with Gaussian initial and observed densities and exploit the analytical results of Section 4 to numerically demonstrate that while two different designs may be approximately equivalent in terms of informing model input parameters, they may perform very differently in terms of informing model predictions. While this may seem fairly obvious to some readers, we find it useful to provide a concrete demonstration of this phenomena for this special case before proceeding to the more complicated examples in subsequent sections.

Here, we take $\Lambda = \mathbb{R}^4$ with $\pi_\Lambda^{\text{init}} \sim N(0, \Sigma_\Lambda)$ where $\Sigma_\Lambda = 2\mathbb{I}_4$. We consider two designs, $D^{(1)} : \Lambda \to \mathbb{R}^2$ and $D^{(2)} : \Lambda \to \mathbb{R}^2$, where

$$D^{(1)} = \begin{bmatrix} -2 & 0 & 0 & 1 \\ 1 & 1 & 0 & -1 \end{bmatrix}, \quad D^{(2)} = \begin{bmatrix} 1 & 0 & 1 & -1 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

In either case, we assume the observed distribution on the data is given by $\pi_{\mathcal{D}}^{\text{obs}} \sim N(0, \Sigma_{\mathcal{D}})$ where $\Sigma_{\mathcal{D}} = 0.8\mathbb{I}_2$. For simplicity of notation, we let $\Sigma_\Lambda^{\text{up},i}$ denote the covariance of the updated density obtained by solving the stochastic inverse problem using the map $D^{(i)}$.

Given these assumptions, we generate 1E6 samples from the initial density to estimate the information gains for the input parameters for both maps as

$$\text{KL}(\pi_\Lambda^{\text{up}} \| \pi_\Lambda^{\text{init}}; D^{(1)}) \approx 1.1140, \text{ and } \text{KL}(\pi_\Lambda^{\text{up}} \| \pi_\Lambda^{\text{init}}; D^{(2)}) \approx 1.0000.$$

Thus, we might consider these two maps to be approximately equally informative with respect to this criteria. Given the analytical expressions for the covariance of the updated densities from Section 4, we can easily compute other OED criteria, such as the determinant of the Hessian (D-optimality),

$$\det\left(\left(\Sigma_\Lambda^{\text{up},1}\right)^{-1}\right) = 2.3437, \quad \det\left(\left(\Sigma_\Lambda^{\text{up},2}\right)^{-1}\right) = 2.3437,$$

the trace of the Hessian (A-optimality),

$$\text{trace}\left(\left(\Sigma_\Lambda^{\text{up},1}\right)^{-1}\right) = 5.0667, \quad \text{trace}\left(\left(\Sigma_\Lambda^{\text{up},2}\right)^{-1}\right) = 4.6667,$$

and the minimum eigenvalue of the Hessian (E-optimality),

$$\min\left(\text{eig}\left(\left(\Sigma_\Lambda^{\text{up},1}\right)^{-1}\right)\right) = 0.5000, \quad \min\left(\text{eig}\left(\left(\Sigma_\Lambda^{\text{up},2}\right)^{-1}\right)\right) = 0.5000.$$

In any case, we find that $D^{(1)}$ and $D^{(2)}$ are approximately equally informative for the model input parameters. However, our conclusions become very different if we consider a linear prediction map, $Q : \Lambda \to \mathbb{R}^2$, given by

$$Q = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 2 \end{bmatrix}.$$

We use rejection sampling to generate samples from each updated density and estimate the KL-divergence between the initial prediction, $\pi_{\mathcal{Q}}^{Q(\text{init})}$ and the updated prediction, $\pi_{\mathcal{Q}}^{Q(\text{up})}$:

$$\text{KL}(\pi_{\mathcal{Q}}^{Q(\text{up})} \| \pi_{\mathcal{Q}}^{Q(\text{init})}; D^{(1)}) \approx 9.8271\text{E-}4, \quad \text{KL}(\pi_{\mathcal{Q}}^{Q(\text{up})} \| \pi_{\mathcal{Q}}^{Q(\text{init})}; D^{(2)}) \approx 0.2406.$$

In fact, for this problem we can analytically determine that $Q\Sigma_\Lambda D^{(1)}$ is a matrix of zeros. Thus, $D^{(2)}$ is much more informative than $D^{(1)}$ with respect to the model predictions despite the fact that both maps are approximately equally informative with respect to the model input parameters (for the metrics considered).

6.2. Linear Elasticity Stress Prediction

As a more realistic example, consider the following computational mechanics problem where a 2-dimensional beam with multiple inclusions is subjected to a uniform loading along the top of the beam (see Figure 2 for an illustration). The computational domain is $\Omega = (0, 6) \times (0, 1)$ and we let Ω_1 and Ω_2 denote the inclusions given by ellipses centered at (2,0.5) and (4,0.5) respectively. Also let $\Omega_0 = \Omega \backslash (\Omega_1 \cup \Omega_2)$ denote the portion of

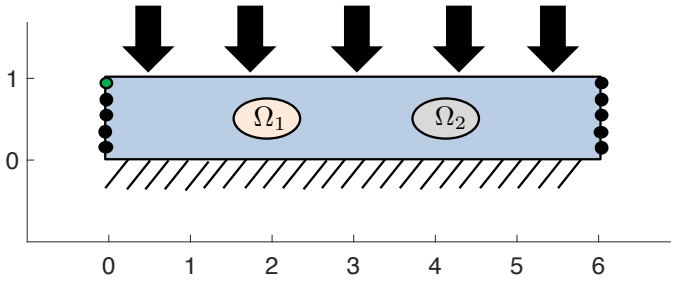

Figure 2: Computational domain for the linear elasticity stress prediction problem. The beam is held fixed along the bottom and a uniform loading is applied along the top. Sensors on each side measure the horizontal displacement.

the beam outside of the inclusions. We use a plane-stress formulation to model displacement and stress and approximate the solution to the governing equations using a standard finite element approximation on a uniform (240×40) quadrilateral mesh. We assume that Young's modulus, E, and the Poisson ratio, ν, are known in Ω_0 but unknown within Ω_1 and Ω_2. Within Ω_0 we set $E_0 = 1.0\text{E}3$ and $\nu_0 = 0.3$ and within each inclusion we assume $E_i \sim U(1\text{E}3, 4\text{E}3)$ and $\nu_i \sim U(0.4, 0.45)$ for $i = 1, 2$, i.e., the inclusions are stiffer than the bulk of the beam and each may be composed of different material. The magnitude of the displacement field and the von Mises stress field for a nominal realization of the parameters are shown in Figure 3.

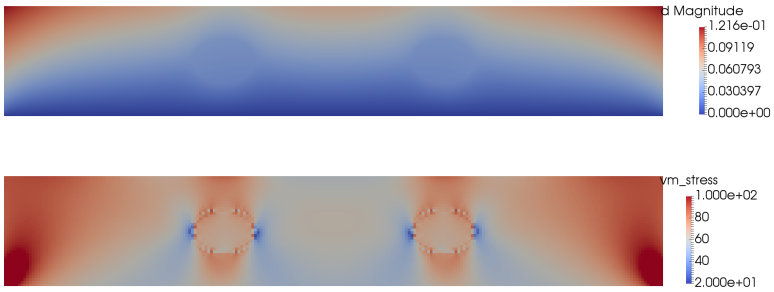

Figure 3: Magnitude of the displacement field (top) and the von Mises stress field (bottom) corresponding to the nominal parameters.

We instrument the model with sensors measuring horizontal displacement at the midpoint of the faces of the elements along the left and right boundaries of the domain. This gives a total of 80 candidate designs. The quantity of interest to predict is the von Mises stress at the center of the right inclusion. This choice is merely for the sake of demonstration and other choices are certainly possible and would presumably give different optimal designs.

To solve the OED problem we must estimate the push-forward of the initial density which we do using 10,000 evaluations of the model with a standard Gaussian kernel density estimator. We must also define a reasonable space for the candidate observed distributions to compute the EIGs of each candidate design. Based on an inspection of the sample-based approximation of the push-forward of the initial density, we fix the standard deviation of the observed density at 1E-5 and assume the mean of the observed density is uniformly distributed over an interval within $\mathcal{D}$. For each

candidate design, $\mathcal{D}^{(k)}$, we choose this interval to be centered at the sample-based approximation of the mean of the push-forward of the initial, which we denote $\overline{\boldsymbol{d}}^{(k)}$, and to have a width equal to 0.5 times the sample-based approximation of the standard deviation of the push-forward of the initial, which we denote $\sigma^{(k)}$. Using the notation of Section 5, we define

$$\mathcal{H} = [\overline{\boldsymbol{d}}^{(k)} - 0.5\sigma^{(k)}, \overline{\boldsymbol{d}}^{(k)} + 0.5\sigma^{(k)}], \quad \mathcal{P}(\boldsymbol{h}) = \{\mathcal{N}(\boldsymbol{h}, \text{1E-10}) : \boldsymbol{h} \in \mathcal{H}\},$$

and define $\mathbb{P}_{\mathcal{H}}$ to be a uniform distribution on $\mathcal{H}$. We also explored different representations of the candidate design space, but the results were comparable and are not presented here.

In Figure 4 (left), we plot the EIG for parameters for each of the candidate designs using 100 random realizations of the parameterization of the observed density. The symmetry in the model produces nearly identical re-

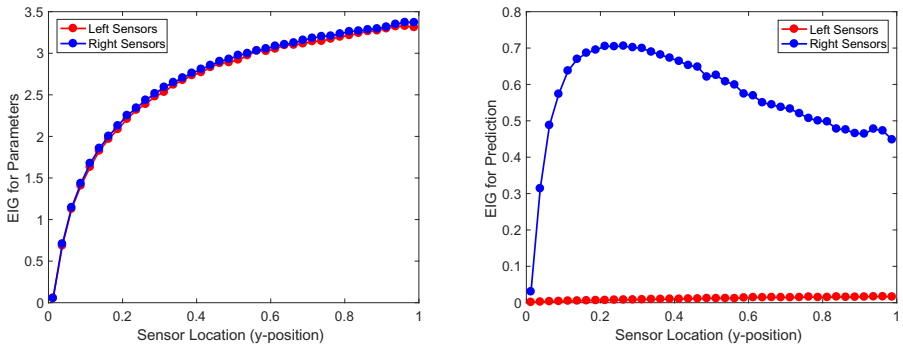

Figure 4: Expected information gains for parameters (left) and for prediction (right).

sults for the sensors on the left and right boundaries. In Figure 4 (right), we show the EIG4P for each of the candidate designs. We see that the sensors along the right boundary, which is closer to the inclusion containing the prediction sensor, are much more informative than the sensors along the left boundary. Thus, the horizontal displacement measured at the sensors on the left and right sides primarily inform the material properties within the left and right inclusions, respectively. This explains the symmetry in the EIG for parameters and the significant difference in the EIG4P. Interestingly, the OED corresponds to measuring the horizontal displacement at the top

of the beam on either side, while the OED4P corresponds to measuring the horizontal displacement at approximately one-fifth of the way up the beam on the right side of the beam. Finally, we note that the EIG in Figure 4 (left) has a significantly larger range than the EIG4P (right). This indicates that the data is informing certain model input parameters more than the model predictions. We have not investigated if this is always the case, but a similar observation is made in the subsequent examples. Of course, this does not affect the OED or OED4P since these are simply different optimization functions.

6.3. Predator Prey Model

In this section, we consider a Lotka-Volterra predator prey model given by,

$$\frac{dx}{dt} = \lambda_1 x - \lambda_2 xy, \tag{16}$$

$$\frac{dy}{dt} = \lambda_3 xy - \lambda_4 y, \tag{17}$$

where x is the number of prey, e.g., mice, and y is the number of predators, e.g., endangered owls. We use a 4^{th}-order explicit Runge-Kutta time integrator with $\Delta t = 0.01$. We assume that the initial conditions are known, but the model parameters, λ_i, are not known precisely. The nominal values for these parameters are 1.0, 0.01, 0.02 and 1.0, respectively, and we assume the parameters are uniformly distributed as follows: $\lambda_1 \sim U(0.95, 1.05)$, $\lambda_2 \sim U(0.01, 0.03)$, $\lambda_3 \sim U(0.02, 0.04)$ and $\lambda_4 \sim U(0.95, 1.05)$. The numerical approximations of the predator and prey populations using the nominal values of these parameters are given in Figure 5 (left). On the right plot of Figure 5, we plot 100 realization of the populations computed from samples from these distributions.

For this demonstration, the prediction QoI is the predator population at $t = 50$, however we assume that direct measurements of the predator population are infeasible or impractical. Thus, we will measure the prey population at an earlier time and hope to inform the model to maximize our information gained in the prediction of the predator population at $t = 50$. We assume that we can measure the prey population at $t = 1.1, 1.2, \ldots, 31.0$ which gives 300 experimental designs. We estimate the push-forward of the initial using 1E4 model evaluations and a standard Gaussian kernel density estimator.

As in Section 6.2, we define the space of candidate observed densities as a family of normal distributions with a fixed variance and a uniformly distributed mean. Based on an inspection of the sample-based approximation

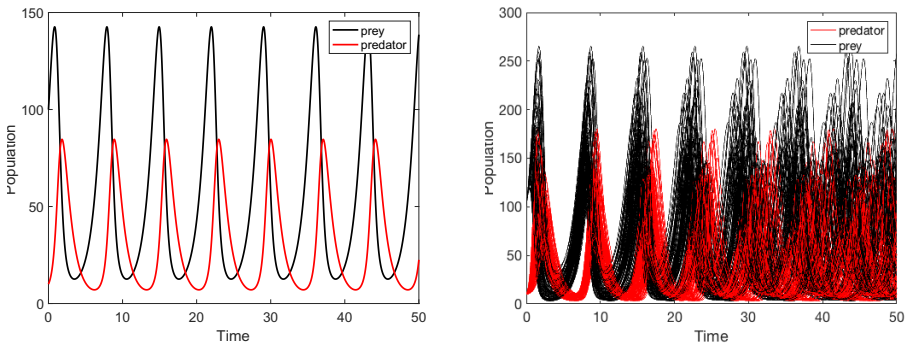

Figure 5: On the left, the numerical approximations of the predator and prey populations using the nominal values of the parameters. On the right, 100 realization of the populations computed from samples from these distributions.

of the push-forward of the initial density, we fix the standard deviation of the observed density at 1.0 and assume the mean of the observed density is uniformly distributed over an interval within $\mathcal{D}$. For each candidate design, $D^{(k)}$, we choose this interval to be centered at the sample-based approximation of the mean of the push-forward of the initial, which we denote $\overline{\boldsymbol{d}}^{(k)}$, and to have a width equal to 0.25 times the sample-based approximation of the standard deviation of the push-forward of the initial, which we denote $\sigma^{(k)}$. The parameters for this normal distribution of the mean are chosen following the same process described in Section 6.2. To be precise, for each candidate design, $\mathcal{D}^{(k)}$, we define

$$\mathcal{H} = [\overline{\boldsymbol{d}}^{(k)} - 0.25\sigma^{(k)}, \overline{\boldsymbol{d}}^{(k)} + 0.25\sigma^{(k)}], \quad \mathcal{P}(\boldsymbol{h}) = \{\mathcal{N}(\boldsymbol{h}, 1.0) : \boldsymbol{h} \in \mathcal{H}\},$$

and define $\mathbb{P}_{\mathcal{H}}$ to be a uniform distribution on $\mathcal{H}$.

In Figure 6 (left) we compare the EIG and the EIG4P for the candidate designs. In Figure 6 (left) we see that while both EIG and EIG4P have periodic structures, the peaks are not aligned. The best design, for the goal of informing model parameters, occurs where EIG is maximized. Similarly the best design for informing predictions maximizes the EIG4P criterion. Based on the data we have, the OED for parameters actually occurs when we measure at $t = 1.6$ and does not provide any significant information gained for the prediction. On the other hand, if we measure at $t = 21.2$

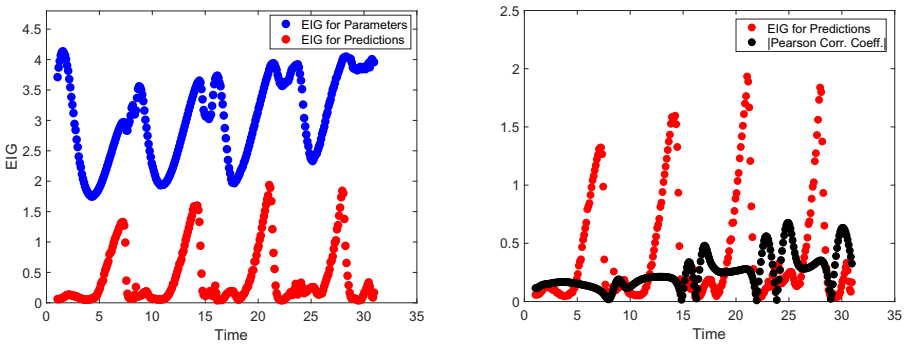

Figure 6: On the left, a comparison of the expected information gained for parameters and for prediction. On the right, a comparison of the expected information gained for predictions and correlation between predator and prey populations.

or $t = 28.0$ then we approximately maximize both the EIG and EIG4P. To select the best design we simply enumerated over all candidate designs. In some situations, it is more computationally efficient to use gradient-based optimization. However, this assumes that the designs can be parameterized by continuous variables. Figure 6 clearly indicates that there are multiple measurement times that provide approximately equivalent local maxima for the EIG. Consequently, using any gradient-based scheme for global optimization of this problem is challenging. In other problems, it may very well be intractable.

In Figure 6 (right) we compare the EIG4P with the absolute value of the Pearson correlation coefficient (PCC) between the prediction QoI (the predator population at $t = 50$) and the observable QoI (the prey populations at earlier times) where the PCC between two random variables, X and Y, is

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y},$$

where σ_X and σ_Y are the standard deviations of X and Y respectively. We see that the PCC and the EIG4P peak at different times. Thus, the optimal time to measure the prey population to reduce the uncertainty in the predator population at a later time ($t = 50$) is not necessarily when these quantities are highly correlated.

6.4. Steady Fluid Flow

In this section, we consider steady-state flow in a channel with multiple obstacles and multiple predictions. The computational domain, shown in Figure 7, is $\Omega = (-5, 10) \times (-3, 3)$. The flow is given by the steady-state

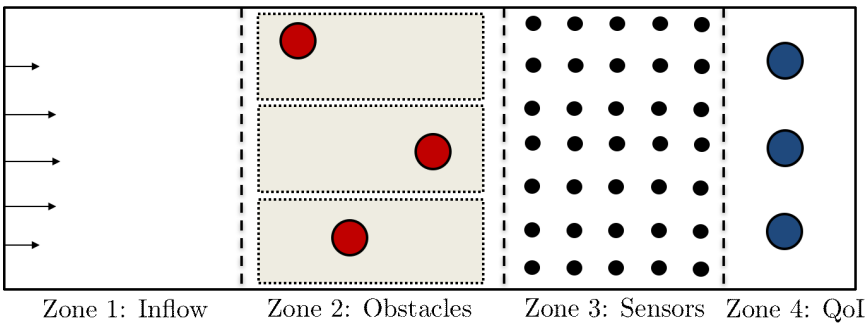

Figure 7: Illustration of the computational domain and zones for steady fluid flow example. This image is just for illustrative purposes and the scales are approximated.

Navier-Stokes equations,

$$\begin{aligned} -\Delta \boldsymbol{u} + Re\boldsymbol{u} \cdot \nabla \boldsymbol{u} + \nabla p &= \mathbf{0}, \\ \nabla \cdot \boldsymbol{u} &= 0. \end{aligned}$$

The flow enters the left boundary with a parabolic profile for the x-velocity. The maximum of this parabolic profile is chosen to give a Reynolds number, Re, of approximately 100, which easily admits a steady solution. The right boundary is an outflow (homogeneous Neumann) boundary and the upper and lower boundaries are no-slip ($\boldsymbol{u} = \mathbf{0}$). The computational domain is decomposed into four regions. In the first region, the flow primarily conforms to the inflow profile. In the second region are three obstacles. We assume that we know the size and shape of these obstacles, but not their precise location. We assume that each obstacle is contained within a rectangular region as shown in Figure 7. Furthermore, we assume that we cannot measure their locations directly. Instead, we can only measure the magnitude of the flow velocity in the wake of the objects (zone 3 in Figure 7). In the fourth zone are three additional objects. We know the location of these objects and our goal is to predict the drop in the pressure from the front to the back of each object.

To summarize, we have 6 independent uncertain parameters that determine the position of the obstacles in zone 2. The initial density is a product

of independent uniform distribution over their respective ranges. There are 600 sensors in zone 3 that measure the magnitude of the flow velocity. Finally, there are three predictions corresponding to the pressure drop for each obstacle in zone 4.

We approximate the solution to the governing equations using a stabilized finite element method on a triangular mesh with continuous piecewise linear approximations for the x-velocity, y-velocity and pressure. For each realization of the location of the obstacles in zone 2 we construct a new mesh. Each of these meshes has approximately 3200 elements giving approximately 5300 degrees-of-freedom, and requires about 11 seconds to generate using Distmesh [28]. Two representative realizations of the obstacle locations and the corresponding meshes are shown in Figure 8. We generate 1000 sam-

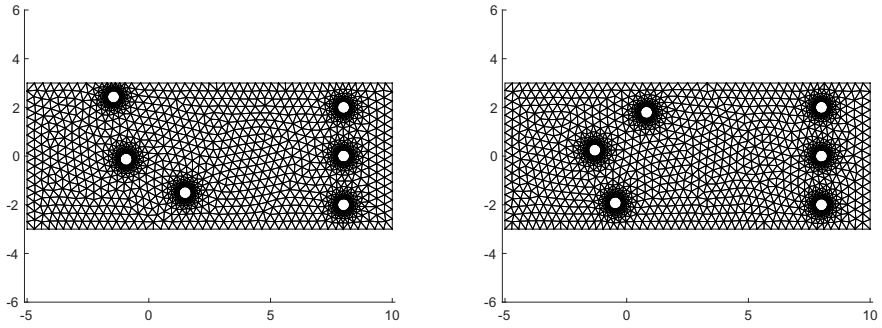

Figure 8: Two random realizations of the obstacle locations and the corresponding finite element mesh.

ples from the initial density and each numerical approximation requires 5-6 Newton steps and the total solution time (not counting mesh construction) is approximately 1.7 seconds using a MATLAB implementation.

We define the space of candidate observed densities as a family of normal distributions with randomly distributed mean and variance. For each candidate design, $D^{(k)}$, we choose this interval to be centered at the sample-based approximation of the mean of the push-forward of the initial, which we again denote $\overline{\boldsymbol{d}}^{(k)}$, and to have a width equal to 0.25 times the sample-based approximation of the standard deviation of the push-forward of the initial, which we again denote $\sigma^{(k)}$. Based on an inspection of the samples

from the push-forward of the initial, we define

$$\mathcal{H}_\sigma := [1.75\text{E-}2, 1.25\text{E-}2], \quad \mathcal{H}_{\overline{d}} := [\overline{\boldsymbol{d}}^{(k)} - 0.25\sigma^{(k)}, \overline{\boldsymbol{d}}^{(k)} + 0.25\sigma^{(k)}].$$

Then we set

$$\mathcal{H} = \mathcal{H}_{\overline{d}} \times \mathcal{H}_\sigma, \quad \mathcal{P}(\boldsymbol{h}) = \left\{\mathcal{N}(h_{\overline{d}}, h_\sigma^2) : h_{\overline{d}} \in \mathcal{H}_{\overline{d}}, h_\sigma \in \mathcal{H}_\sigma\right\},$$

and define $\mathbb{P}_{\mathcal{H}}$ to be a uniform distribution on $\mathcal{H}$.

We plot the EIG in Figure 9. We see that if the goal is to determine

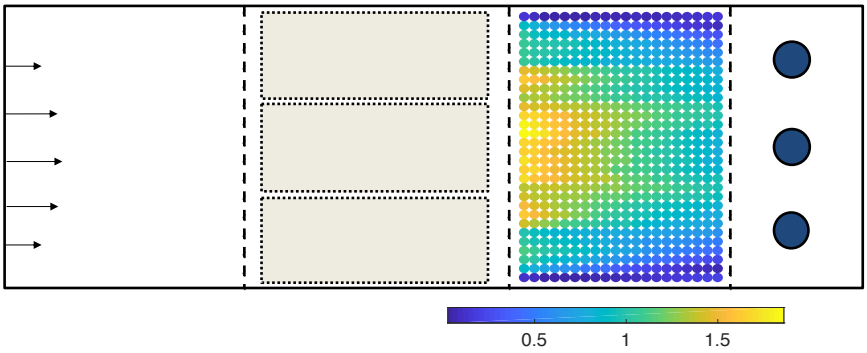

Figure 9: Expected information gained for parameters for the stead-state fluid flow model.

the OED for parameters, i.e., to maximize the EIG to locate the obstacles, the optimal location to place the sensor is close to the obstacles and in the middle (in the y-direction) of the domain. We contrast this with the EIG4P shown in Figure 10. If the goal is to determine the OED4P, i.e., to maximize the EIG4P to predict the pressure drop for each of the obstacles in zone 4, then the optimal location for a sensor is directly upstream of one of the obstacles.

For this model, both the OED and OED4P are consistent with our physical intuition for steady-state fluid flow. Moreover, as with the previous example, the EIG and EIG4P criteria both have several local maxima. This hinders the ability of any gradient-based optimization procedure to identify a global optima given the strong dependence of such procedures on their initialization.

7. Conclusion

This paper presents an approach for optimal experimental design (OED) that selects experimental data which significantly influence uncertainty in

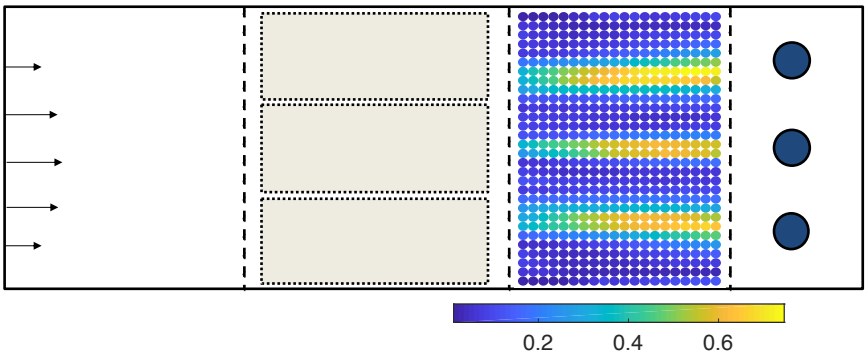

Figure 10: Expected information gained for predictions for the stead-state fluid flow model.

model predictions that cannot be observed directly. This is in contrast to most traditional OED strategies which focus on reducing uncertainty in estimates of model parameters. The method presented, for optimal experimental design for prediction (OED4P), utilizes a recently developed framework for stochastic inversion based on push-forward probability measures which, under certain assumptions on the problem formulation, facilitates a computationally efficient and scalable approach for OED4P. Traditional OED methods, which are often formulated as an inverse problem wrapped within an optimization procedure, often requiring thousands or millions of model evaluations. In contrast, the proposed method for OED4P only requires solving a single forward UQ problem under certain restrictions on the design space.

The difference between a traditional form of OED and OED4P is demonstrated on a number of numerical examples. The results highlight that when prediction is the ultimate modeling objective, experimental designs that target reduction in parameter uncertainty may be inefficient, or worse yet, entirely ineffective for prediction.

The primary goal of this paper is to motivate the need for OED4P. We focused on computing designs which consists of only a limited number of experiments. Such situations often arise when performing field or laboratory experiments is extremely expensive. The computational cost of the approach presented here grows rapidly with the number of potential experiments. Future work will look to exploit the connections with deterministic optimization, established in [33], to develop scalable approaches for OED4P which can generate experimental designs with large numbers of experiments. We are also pursuing alternative approaches that avoid the use of kernel den-

sity approximations in the output space to enable the proposed methodology to be extended to high-dimensional observed and/or prediction spaces.

8. Acknowledgments

T. Wildey's work was supported by the U.S. Department of Energy, Office of Science, Early Career Research Program. T. Butler's work was supported by the National Science Foundation under Grant No. (DMS-1818941). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. J. Jakeman's work was partially supported by DARPA EQUIPS.

[1] Alexanderian, A., Petra, N., Stadler, G., and Ghattas, O. A-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems with regularized ℓ_0-sparsification. *SIAM J. Sci. Comput. 36*, 5 (2014), A2122–A2148.

[2] Alexanderian, A., Petra, N., Stadler, G., and Ghattas, O. A fast and scalable method for A-optimal design of experiments for infinite-dimensional Bayesian nonlinear inverse problems. *SIAM J. Sci. Comput. 38*, 1 (2016), A243–A272.

[3] Alexanderian, A., and Saibaba, A. Efficient D-optimal design of experiments for infinite-dimensional Bayesian linear inverse problems. *SIAM J. Sci. Comput. 40*, 5 (2018), A2956–A2985.

[4] Atkinson, A., and Donev, A. *Optimum Experimental Designs*. Oxford University Press, 1992.

[5] Attia, A., Alexanderian, A., and Saibaba, A. K. Goal-oriented optimal design of experiments for large-scale Bayesian linear inverse problems. *Inverse Problems 34*, 9 (jul 2018), 095009.

[6] Bauer, I., Bock, H. G., Krkel, S., and Schlder, J. P. Numerical methods for optimum experimental design in DAE systems. *J. Comput. Appl. Math. 120*, 12 (2000), 1 – 25.

[7] Bock, H. G., Körkel, S., and Schlöder, J. P. *Parameter Estimation and Optimum Experimental Design for Differential Equation Models*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 1–30.

[8] Butler, T., Jakeman, J., Pisolov, M., Walsh, S., and Wildey, T. A new approach to optimal experimental design using singular values of Jacobians for non-linear observable maps. *Comput. Meth. Appl. Mech. Engrg.* (2018). Submitted, arXiv:1811.04988.

[9] Butler, T., Jakeman, J., and Wildey, T. Combining push-forward measures and Bayes rule to construct consistent solutions to stochastic inverse problems. *SIAM J. Sci. Comput. 40*, 2 (2018), A984–A1011.

[10] Butler, T., Jakeman, J., and Wildey, T. Convergence of probability densities using approximate models for forward and inverse problems in uncertainty quantification. *SIAM J. Sci. Comput. 40*, 5 (2018), A3523–A3548.

[11] Chaloner, K., and Verdinelli, I. Bayesian experimental design: A review. *Statist. Sci. 10*, 3 (08 1995), 273–304.

[12] Cover, T. A., and Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons, 2006.

[13] Cox, D. R., and Reid, N. *The theory of the design of experiments*. CRC Press, 2000.

[14] Dellacherie, C., and Meyer, P. *Probabilities and Potential*. North-Holland Publishing Co., Amsterdam, 1978.

[15] Fisher, R. A. *The Design of Experiments*, 8th ed. Oliver & Boyd, Edinburgh, United Kingdom, 1966.

[16] Haber, E., Horesh, L., and Tenorio, L. Numerical methods for the design of large-scale nonlinear discrete ill-posed inverse problems. *Inverse Problems 26*, 2 (2010), 025002.

[17] Haber, E., Magnant, Z., Lucero, C., and Tenorio, L. Numerical methods for A-optimal designs with a sparsity constraint for ill-posed inverse problems. *Computational Optimization and Applications 52*, 1 (2012), 293–314.

[18] Horesh, L., Haber, E., and Tenorio, L. *Optimal Experimental Design for the Large-Scale Nonlinear Ill-Posed Problem of Impedance Imaging*. John Wiley & Sons, Ltd, 2010, pp. 273–290.

[19] Huan, X., and Marzouk, Y. M. Simulation-based optimal Bayesian experimental design for nonlinear systems. *J. Comput. Phys. 232*, 1 (2013), 288 – 317.

[20] Lieberman, C., and Willcox, K. Goal-oriented inference: Approach, linear theory, and application to advection diffusion. *SIAM J. Sci. Comput. 34*, 4 (2012), A1880–A1904.

[21] Lieberman, C., and Willcox, K. Nonlinear goal-oriented Bayesian inference: Application to carbon capture and storage. *SIAM J. Sci. Comput. 36*, 3 (2014), B427–B449.

[22] Long, Q., Scavino, M., Tempone, R., and Wang, S. Fast estimation of expected information gains for Bayesian experimental designs based on laplace approximations. *Comput. Meth. Appl. Mech. Engrg. 259* (2013), 24 – 39.

[23] Long, Q., Scavino, M., Tempone, R., and Wang, S. A Laplace method for under-determined Bayesian optimal experimental designs. *Comput. Meth. Appl. Mech. Engrg. 285* (2015), 849–876.

[24] Loredo, T. J., and Chernoff, D. F. *Bayesian Adaptive Exploration.* Springer New York, New York, NY, 2003, pp. 57–70.

[25] Marvin, B., Bui-Thanh, T., and Wildey, T. A scalable approach for solving stochastic inverse problems based on push-forward measures and Bayes rule. In preparation, 2019.

[26] Müller, P., Sansó, B., and Iorio, M. D. Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association 99*, 467 (2004), 788–798.

[27] Pazman, A. *Foundations of Optimum Experimental Design.* D. Reidel Publishing Co., 1986.

[28] Persson, P., and Strang, G. A simple mesh generator in matlab. *SIAM Review 46*, 2 (2004), 329–345.

[29] Pukelsheim, F. *Optimal Design of Experiments.* John Wiley & Sons, New-York, 1993.

[30] Solonen, A., Haario, H., and Laine, M. Simulation-based optimal design using a response variance criterion. *Journal of Computational and Graphical Statistics 21*, 1 (2012), 234–252.

[31] Ucinski, D. *Optimal measurement methods for distributed parameter system identification.* CRC Press, Boca Raton, 2005.

[32] van Erven, T., and Harremoes, P. Renyi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory 60* (2014), 3797–3820.

[33] Walsh, S., Wildey, T., and Jakeman, J. Optimal experimental design using a consistent Bayesian approach. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering 4*, 1 (2017), 1–19.